Vol.2010-SLP-80 No.8
2010/2/13

最小相対エントロピー識別学習へのラティスによる仮説表現と並列化可能な最適化手法の導入

久保 陽太郎[†1] 渡部 晋治[†2]
中村 篤[†2] 小林 哲則[†1]

識別学習は，デコーダの出力する認識仮説と比較して正解ラベルの尤度を相対的に高めることで識別に特化したモデルを得るための手法であるが，経験的に過学習しやすいことが知られている．近年，音響モデルの識別学習において過学習を軽減するため，最小相対エントロピー識別が音響モデルの識別学習に導入されてきた．この手法ではパラメタ推定の不確実性をパラメタ分布によって表現することで適切に取り扱うことを可能としており過学習に強いと考えられるが，従来の実現法では大量の認識仮説，および大量のトレーニングデータを取り扱うには膨大な量の計算を単一のコンピュータで実行しなければならなかった．そこで，本研究では，ラティス型認識仮説表現を導入することで認識仮説の数に対する計算効率を，また勾配法に基づく並列化可能な最適化法を導入することでトレーニングデータの数に対する並列計算効率を向上させた．提案法を用いることで，最小相対エントロピー識別学習に必要なステップのほぼ全てがグリッドコンピュータのような並列計算環境で実現可能になり，また，従来の N-best に基づく認識仮説表現では表現しきれないような膨大な数の認識仮説に対する最適化が行なえるようになった．

Parallelizable Optimization Methods and Lattice-based Representations for Minimum Relative Entropy Discrimination Training

Yotaro Kubo,[†1] Shinji Watanabe,[†2]
Atsushi Nakamura[†2] and Tetsunori Kobayashi [†1]

In order to improve the performance of automatic speech recognition, discriminative training methods are introduced for training processes of acoustic models in speech recognizers. Recently, minimum relative entropy discrimination (MRED) training of acoustic models is introduced in order to prevent overfitting problems in discriminative training methods by representing parameters as random variables. Despite of these advantages, the conventional implementation of MRED lacks scalability to the amount of training dataset and the number of the hypothesis label sequences obtained from decoders. In this study, we attempt to improve scalability of MRED training. The lattice-based representations of the hypothesis label sequences are introduced in order to improve scalability due to the number of the hypothesis label sequences. Further, the gradient-based optimization method is introduced in order to ensure parallelism in the MRED training method. By incorpolating proposed methods, it is confirmed that the MRED training procedure can now be performed in parallel computing environments such as grid computers. Furthremore, the large number of the hypothesis label sequences can be handled in the MRED by using hypothesis lattices obtained from decoders.

1. はじめに

近年，音響モデルの学習／推定問題に対し，識別学習と呼ばれる一連の手法群が成果を挙げることが実験的に明らかになってきている．識別学習手法では，音声特徴量系列の精密な生成モデルを得ることを目的とした最尤推定法やベイズ推論法と異なり，識別性能を最大にすることのできるモデルを得ることを目的とする．しかし，識別学習は経験的に過学習しやすいことが知られており，その効果を充分に得るためには，大量のトレーニングデータを用いて学習を行なう必要がある．

加えて，識別学習に利用される認識仮説の数も識別学習の効果を決定する重要な要素である．識別学習は一般に，デコーダが内部的に用いている認識仮説群を利用し，正解系列の尤度を認識仮説中において相対的に高めることで達成される．従来より仮説群の表現には，N-best に基づく手法 [1] や，ラティスに基づく手法 [2] が利用されてきた．

Jebara らによって提案された最小相対エントロピー識別学習 [3] は，ベイズ推論と同様に，パラメタの分布を陽に考慮することにより，過学習の問題を緩和する識別モデルの一手法であり，音声認識における HMM の識別学習にも応用されてきた [4]．しかし，たとえ過学習の問題が緩和されたとしても，大量のデータおよび認識仮説群を扱うことは依然として必要であり，これらの増加に対するスケーラビリティを保証することは重要な課題である．

本稿では，最小相対エントロピーの大規模データについてのスケーラビリティを向上させるための基礎検討として，切除平面法にヒントを得たラティス型の対立候補の表現を導入し，識別関数近似精度の向上と，対立候補表現のメモリ効率向上の両方を同時に達成することを試みる．加えて，計算量についてのスケーラビリティを確保するため，勾配法をベースとした最適化法についても検討を行なう．従来，MRED におけるラグランジュ未定乗数の最適化では，SVM 等の凸最適化でよく用いられる分割統治法を用いて実現されてきた [4]．分割統治法は効率の良い最適化が可能であることが知られているが，各ラグランジュ未定乗数の 1 ステップ分の更新が，別のラグランジュ未定乗数の更新に影響を与えるため，並列化が非常に難しい．本稿では，凸最適化由来の最適化法ではなく，より一般的な最適化技法である勾配法を用いて最適化を実現することで，並列化可能なアルゴリズムを導出する．

†1 早稲田大学
Waseda University
†2 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

2. ラティスによる仮説表現に基づく最小相対エントロピー識別学習

本稿ではトレーニングデータセットに含まれる i 番目の特徴量系列を $\boldsymbol{X}^{(i)}$，対応する正解ラベル系列を $\boldsymbol{l}^{(i)}$ と置く．ここで $\boldsymbol{X}^{(i)}$ は D 次元ベクトルの系列であり，系列中の各フレームは $\boldsymbol{x}_n^{(i)} \in \mathbb{R}^D$ $(n \in [1..N^{(i)}])$ で表わす．

2.1 連続分布型 HMM (CD-HMM) の最小相対エントロピー識別学習

最小相対エントロピー識別ではパラメタ Θ の分布表現 $P(\Theta)$ を陽に導入するため，同じ識別性能を持つパラメタ分布の中で，最も事前分布 $P^{(0)}(\Theta)$ に近い分布を事後分布 $P(\Theta)$ として定義することを考える．事前分布とのダイバージェンスを KL ダイバージェンス (相対エントロピー) によって表現し，識別性能を識別関数 $\mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta)$ に関する閾値 $\xi^{(i)}$(スラック変数) によって表現することで，以下の最適化問題の解としてパラメタとスラック変数の事後分布 $P(\Theta, \xi)$ を得ることを考える．

$$\underset{P(\Theta,\xi)}{\text{minimize}}\ \mathrm{KL}\left[P(\Theta,\xi)||P^{(0)}(\Theta,\xi)\right], \text{subject to } \left\langle \mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta) - \xi^{(i)} \right\rangle_{P(\Theta,\xi)} \geq 0 \quad \forall i \tag{1}$$

ここで，$\langle f(x) \rangle_{g(x)}$ は変数 x が確率分布関数 $g(x)$ に従うときの関数 $f(x)$ の期待値であり，$\langle f(x) \rangle_{g(x)} = \int_x g(x) f(x) dx$ と定義される．また，$\mathrm{KL}[p(x)||q(x)]$ は確率分布関数 p から q への KL ダイバージェンスであり，$\langle \log p(x) - \log q(x) \rangle_{p(x)}$ と定義される．

凸最適化問題ではラグランジュ汎関数と呼ばれる関数を導入し，その鞍点を求めることで解の別表現を得ることができる [9]．これを利用することで，上記問題 (式 (1)) の最適解はラグランジュ未定乗数と呼ばれる非負変数 $\alpha^{(i)} \geq 0$ を導入した上で，以下のように示されることがわかる．

$$P(\Theta, \xi) \propto P^{(0)}(\Theta, \xi) \exp\left\{ \sum_i \alpha^{(i)} \left(\mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta) - \xi^{(i)} \right) \right\} \tag{2}$$

最小相対エントロピー識別では，ラグランジュ未定乗数 $\alpha^{(i)}$ を最適化によって求めることで，事後分布を決定する．

2.2 ラティスを用いた識別関数の定義

識別関数として以下の *Minimum classification error* (MCE) 型の識別関数 (MCE における目的関数) [1] を導入することを考える．[⋆1]

$$\begin{aligned}\mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta) &\stackrel{\mathrm{def}}{=} \log \frac{P(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}|\Theta)}{\max_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta)} = \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}|\Theta) - \log \max_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta) \\ &= \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}|\Theta) - \max_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta)\end{aligned} \tag{3}$$

ここで，最後の等式は log の単調増加性のため成立する．

最小相対エントロピー識別ではラグランジュ未定乗数に関する最適化 $\alpha^{(i)}$ において，識別関数の期待値計算が必要となる．そのため，識別関数中に不連続な関数 max が含まれていると，解析的に目的関数を展開できない．そこで，max を以下の連続関数で近似することを考える．

$$\max_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta) \approx \frac{1}{\eta} \log \sum_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} \eta P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta) \stackrel{\mathrm{def}}{=} \mathrm{softmax}^{\eta}_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta) \tag{4}$$

softmax は $\eta \to \infty$ で max 関数に収束する．以降，解析的に取り扱いが容易な $\mathrm{softmax}^{\eta=1}$ を用いて近似を行なう．式 (4) を識別関数 (式 (3)) に代入し，$\eta = 1$ を代入することで，以下の近似識別関数を得る．

$$\mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta) \approx \log P(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}|\Theta) - \log \sum_{\boldsymbol{l} \neq \boldsymbol{l}^{(i)}} P(\boldsymbol{X}^{(i)}, \boldsymbol{l}|\Theta) \tag{5}$$

ここで正解ラベルに対応する正解ラティス $A^{(i)}$ と正解ラベル以外の候補集合を表わす不正解ラティス $\tilde{A}^{(i)}$ を導入する．連続分布型 HMM の隠れ変数は時刻 n に対応する混合要素インデックス m_n と状態インデックス s_n の系列 $q \stackrel{\mathrm{def}}{=} \{q_n = (m_n, s_n) | \forall n\}$ であると考えることができる．また，ラティスは取り得る隠れ変数 (特に状態系列) を制限するものであると考えることができ，ラティスが与えられた時の確率はラティスから取り得る隠れ変数 q について周辺化したものとして定義できる．この関係を用いて識別関数を書き直すことで以下を得る．

$$\begin{aligned}\mathcal{D}(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)}; \Theta) &\approx \log \sum_{q \in \mathcal{Q}(A^{(i)})} P(\boldsymbol{X}^{(i)}, q|\Theta) - \log \sum_{q \in \mathcal{Q}(\tilde{A}^{(i)})} P(\boldsymbol{X}^{(i)}, q|\Theta) \\ &\stackrel{\mathrm{def}}{=} \mathcal{L}(\boldsymbol{X}^{(i)}, A^{(i)}; \Theta) - \mathcal{L}(\boldsymbol{X}^{(i)}, \tilde{A}^{(i)}; \Theta)\end{aligned} \tag{6}$$

ここで $\mathcal{Q}(A)$ はラティス A から，取り得る隠れ変数の集合を返す関数である．不正解ラティスは一般に認識仮説ラティスから正解ラティスに対応する系列を FST の Difference 演算 [7] を用いて取り除くことで得られる．また，$\mathcal{L}(\boldsymbol{X}, A; \Theta)$ はラティス A と特徴量系列 $\boldsymbol{X}$，モデルパラメタ Θ を受けとり対数尤度を返す関数である．

2.3 識別関数の近似と双対問題の導出

取り得る q に対応する離散確率分布 $\omega = \{\omega_q \geq 0 | \forall q \in \mathcal{Q}(A)\}$ (ここで $0 < \omega_q < 1$ かつ $\sum_{q \in \mathcal{Q}(A)} \omega_q = 1$) を導入し，Jensen の不等式を適用することで，対数尤度関数 $\mathcal{L}$ の下界 $\tilde{\mathcal{L}}$ を以下のように定めることができる．

$$\begin{aligned}\mathcal{L}(X, A; \Theta) &= \log \sum_{q \in \mathcal{Q}(A)} P(\boldsymbol{X}, q|\Theta) = \log \sum_{q \in \mathcal{Q}(A)} \omega_q \frac{P(\boldsymbol{X}, q|\Theta)}{\omega_q} \\ &\geq \sum_{q \in \mathcal{Q}(A)} \omega_q \log P(\boldsymbol{X}, q|\Theta) + H(\omega) \stackrel{\mathrm{def}}{=} \tilde{\mathcal{L}}(X, A; \Theta, \omega)\end{aligned} \tag{7}$$

ここで $H(\omega)$ は離散確率分布 ω のエントロピーである．

$\tilde{\mathcal{L}}$ は対数尤度関数 $\mathcal{L}$ と接するため，以下のような関係を満たす．

$$\mathcal{L}(X, A; \Theta) = \max_{\omega} \tilde{\mathcal{L}}(X, A; \Theta, \omega) \tag{8}$$

この関係性とラティス型識別関数 (式 (6)) を主問題 (式 (1)) に代入すると，以下の最適

⋆1 識別関数は与えられたモデルパラメタ Θ のトレーニングデータ $(\boldsymbol{X}^{(i)}, \boldsymbol{l}^{(i)})$ に関する識別性能を示す関数であれば任意のものが利用可能であり，MCE 型の他に *Maximum mutual information* (MMI) 型 [5] や *Minimum phone error* (MPE) 型 [6] のようなものも考えられる．

Vol.2010-SLP-80 No.8
2010/2/13

化問題を得ることができる．

$$\begin{aligned}&\underset{P(\Theta,\xi)}{\text{minimize}}\ \mathrm{KL}\left[P(\Theta,\xi)||P^{(0)}(\Theta,\xi)\right]\\&\text{subject to}\ \left\langle\max_{\omega}\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega)-\max_{\tilde{\omega}}\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},\tilde{A}^{(i)};\Theta,\tilde{\omega})-\xi^{(i)}\right\rangle_{P(\Theta,\xi)}\geq 0\quad\forall i\end{aligned}\tag{9}$$

subject to 以下にある max の期待値を解析的に取り扱うのは難しいため，ここで，max の期待値を期待値の max で近似し，以下の問題を解くことを考える．

$$\begin{aligned}&\underset{P(\Theta,\xi)}{\text{minimize}}\ \mathrm{KL}\left[P(\Theta,\xi)||P^{(0)}(\Theta,\xi)\right]\\&\text{subject to}\\&\max_{\omega}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega)\right\rangle_{P(\Theta,\xi)}-\max_{\tilde{\omega}}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},\tilde{A}^{(i)};\Theta,\tilde{\omega})\right\rangle_{P(\Theta,\xi)}-\left\langle\xi^{(i)}\right\rangle_{P(\Theta,\xi)}\geq 0\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall i\end{aligned}\tag{10}$$

ここで，形式的に i 番目のトレーニングデータに対して定義できる，全ての $\tilde{\omega}$ の集合 $\tilde{\Omega}^{(i)}$ を導入することで，制約部を以下のような等価な制約に書き直すことができる．

$$\begin{aligned}&\underset{P(\Theta,\xi)}{\text{minimize}}\ \mathrm{KL}\left[P(\Theta,\xi)||P^{(0)}(\Theta,\xi)\right]\\&\text{subject to}\\&\max_{\omega}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega)\right\rangle_{P(\Theta,\xi)}-\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},\tilde{A}^{(i)};\Theta,\tilde{\omega})\right\rangle_{P(\Theta,\xi)}-\left\langle\xi^{(i)}\right\rangle_{P(\Theta,\xi)}\geq 0\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall i\ \forall\tilde{\omega}\in\tilde{\Omega}^{(i)}\end{aligned}\tag{11}$$

この最適化問題は $\tilde{\Omega}^{(i)}$ の要素数に対応する数の制約，すなわち無限個の制約を取り扱う必要があり，計算機上での実現は不可能である．そのため，有限個の要素を持つサブセット $\tilde{\Omega}^{(i)}\overset{\text{def}}{=}\{\tilde{\omega}_1^{(i)},\tilde{\omega}_2^{(i)},\cdots,\tilde{\omega}_R^{(i)}\}$ の上での最適化に緩和することを考える．緩和した後の最適化問題の解は，式 (2) と同様に以下の関数型を持つことがわかる．

$$\begin{aligned}&P(\Theta,\xi)\propto P^{(0)}(\Theta,\xi)\exp\left\{\sum_i\sum_{r=1}^{R}\alpha_r^{(i)}\left(\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega^{(i)})-\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},\tilde{A}^{(i)};\Theta,\tilde{\omega}_r^{(i)})-\xi^{(i)}\right)\right\},\\&\omega^{(i)}=\underset{\omega}{\text{argmax}}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)},\Theta,\omega)\right\rangle_{P(\Theta,\xi)}\end{aligned}\tag{12}$$

ここで，$\alpha\overset{\text{def}}{=}\{\alpha_r^{(i)}|\forall i,1\leq r\leq R\}$ はラグランジュ未定乗数である．式 (2) の場合と異なり，制約を R に対応する数だけ増やしているため，ラグランジュ未定乗数の数も増えている．また，ω は式 (11) 内の制約部に表われる最適化の結果として得られるものである．

この事後分布をラグランジュ汎関数に代入し，その鞍点を直接求める最適化に変形することで，主問題 (式 (11)) に対する双対問題を以下のように得ることができる．

$$\underset{\alpha}{\text{maximize}}\ J(\alpha,\Omega)=-\log Z(\alpha,\Omega)\quad\text{subject to}\quad\alpha_r^{(i)}\geq 0\quad\forall i\ \forall r\tag{13}$$

ここで

$$\begin{aligned}&Z(\alpha,\Omega)=\left\langle\exp\left[\sum_i\sum_{r=1}^{R}\alpha_r^{(i)}\left(\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega^{(i)})-\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},\tilde{A}^{(i)};\Theta,\tilde{\omega}_r^{(i)})-\xi^{(i)}\right)\right]\right\rangle_{P^{(0)}(\Theta,\xi)}\\&\omega^{(i)}=\underset{\omega}{\text{argmax}}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)},\Theta,\omega)\right\rangle_{P(\Theta,\xi)}\end{aligned}\tag{14}$$

ここで Ω は $\omega^{(i)}$ の集合であり，$\Omega\overset{\text{def}}{=}\{\omega^{(i)}|\forall i\}$ である．なお，目的関数 $J(\alpha,\Omega)$ は α について凹 (concave)，Ω については非凹 (非凸) である．

2.4 共役事前分布の導入

この目的関数 (式 (14)) 中の期待値計算は，事前分布 $P^{(0)}(\Theta,\xi)$ として共役な事前分布を用いることで解析的に求めることができる．本研究では，CD-HMM の各パラメタに対して共役な事前分布として以下のものを用いた．

CD-HMM 内で利用される各正規分布には対角共分散性を仮定し，g 番目の正規分布の平均ベクトル $\boldsymbol{\mu}_g$ および精度ベクトル $\boldsymbol{\tau}_g$ の事前分布 $P^{(0)}(\boldsymbol{\mu}_g,\boldsymbol{\tau}_g)$ を，以下のように正規-ガンマ分布を用いて定義した．

$$P^{(0)}(\boldsymbol{\mu}_g,\boldsymbol{\tau}_g)=\prod_{d=1}^{D}\mathcal{N}\circ\mathcal{G}(\mu_{g,d},\tau_{g,d}|\mu_{g,d}^{(0)},\gamma_{g,d}^{(0)},\eta_g^{(0)},R_{g,d}^{(0)})\tag{15}$$

ここで $\mathcal{N}\circ\mathcal{G}$ は正規-ガンマ分布の分布関数であり，以下のように表わされる．

$$\begin{aligned}&\mathcal{N}\circ\mathcal{G}(\mu,\tau|\mu^{(0)},\gamma^{(0)},\eta^{(0)},R^{(0)})\propto\\&\qquad\frac{(R^{(0)})^{\eta^{(0)}}}{\Gamma\left(\eta^{(0)}\right)}(\tau)^{\eta^{(0)}-1/2}\exp\left\{-R^{(0)}\tau-\frac{R^{(0)}\gamma^{(0)}}{2}(\mu^{(0)}-\mu)^2\right\}\end{aligned}\tag{16}$$

ここで $\mu_{g,d}^{(0)},\gamma_{g,d}^{(0)},\eta_g^{(0)},R_{g,d}^{(0)}$ はハイパーパラメタである．

また s 番目の状態に対応する混合分布の重みベクトル $\boldsymbol{\rho}_s\overset{\text{def}}{=}[\rho_{s,1},\rho_{s,2},\cdots]^{\mathsf{T}}$ に対しては，以下の Dirichlet 分布を利用した．

$$P^{(0)}(\boldsymbol{\rho}_s|\boldsymbol{\phi}_s^{(0)})=\frac{\Gamma\left(\sum_m\phi_{s,m}^{(0)}\right)}{\prod_m\Gamma\left(\phi_{s,m}^{(0)}\right)}\prod_m(\rho_{s,m})^{\phi_{s,m}^{(0)}}\tag{17}$$

状態遷移確率行列 $\boldsymbol{\Pi}$ も行毎の事前分布として Dirichlet 分布を用いることで，解析的に目的関数を展開することができるが，紙幅の都合上，本稿では省略する[⋆1]．

また，i 番目の制約に関するスラック変数の事前分布としてはスラック変数が十分に大きい，すなわち識別関数 $\tilde{\mathcal{D}}$ の値が十分に大きくなければ制約が満たされないということを記述するため，以下の指数分布を用いた．

$$P^{(0)}(\xi^{(i)}|C^{(0)},L^{(0)})=\frac{1}{C^{(0)}}\exp\left\{-C^{(0)}\left(\xi^{(i)}-L^{(0)}N^{(i)}\right)\right\},\quad(\xi_r^{(i)}\leq L^{(0)}N^{(i)}).\tag{18}$$

ここで $L^{(0)}$ はどの程度 $\tilde{\mathcal{D}}$ が大きければ十分かを示すハイパーパラメタであり，$C^{(0)}>0.0$ はその値に達していない時のペナルティを示すハイパーパラメタである．

⋆1 後述の実験でも，遷移確率行列を学習対象として扱わず，最尤推定のものをそのまま利用した．

2.5 事後分布関数の具体形

前節で定義した事前分布を式 (12) で示される事後分布関数に代入することで，以下の事後分布関数を得る．

$$P(\Theta,\xi|\alpha,\Omega)=\prod_g P(\boldsymbol{\mu}_g,\boldsymbol{\tau}_g|\alpha,\Omega)\times\prod_s P(\boldsymbol{\rho}_s|\alpha,\Omega)\times\prod_i P(\xi^{(i)}|\alpha,\Omega) \tag{19}$$

ここでは正確性のため，全ての確率分布関数が α と Ω に依存することを明示した．

スラック変数の事後分布は音響モデルの構築には利用されないため，正規分布パラメタの事後分布と，混合重みベクトルの事後分布を考える．正規分布パラメタの事後分布は，以下のように展開される．

$$P(\boldsymbol{\mu}_g,\boldsymbol{\tau}_g|\alpha,\Omega)=\prod_d \mathcal{N}\circ\mathcal{G}(\mu_{g,d},\tau_{g,d}|\hat{\mu}_{g,d}(\alpha,\Omega),\hat{\gamma}_g(\alpha,\Omega),\hat{\eta}_g(\alpha,\Omega),\hat{R}_{g,d}(\alpha,\Omega)) \tag{20}$$

ここで $\hat{\mu}_{g,d},\hat{\gamma}_{g,d},\hat{\eta}_{g,d},\hat{R}_{g,d}$ は α,Ω が与えられた時の事後分布パラメタを示す関数であり，以下の式で示される．

$$\begin{aligned}\hat{\mu}_{g,d}(\alpha,\Omega)&=\frac{\gamma_g^{(0)}\mu_{g,d}^{(0)}+\Delta^1_{g,d}(\alpha,\Omega)}{\gamma_g^{(0)}+\Delta^0_g(\alpha,\Omega)},\quad \hat{\gamma}_g(\alpha,\Omega)=\gamma_g^{(0)}+\Delta^0_g(\alpha,\Omega),\\ \hat{\eta}_g(\alpha,\Omega)&=\eta_g^{(0)}+\tfrac{1}{2}\Delta^0_g(\alpha,\Omega),\quad \hat{R}_{g,d}(\alpha,\Omega)=R_{g,d}^{(0)}+\Delta^2_g(\alpha,\Omega).\end{aligned} \tag{21}$$

本稿では，$\Delta^0_g,\Delta^1_{g,d},\Delta^2_{g,d}$ をそれぞれ，零次／一次／二次の差分統計量と呼び，以下のように定義する．

$$\begin{aligned}\Delta^0_g(\alpha,\Omega)&=\sum_i\sum_{r=1}^R\alpha_r^{(i)}\left(\chi^0_g(\boldsymbol{X}^{(i)},\omega^{(i)})-\chi^0_g(\boldsymbol{X}^{(i)},\tilde{\omega}_r^{(i)})\right),\\ \Delta^1_{g,d}(\alpha,\Omega)&=\sum_i\sum_{r=1}^R\alpha_r^{(i)}\left(\chi^1_{g,d}(\boldsymbol{X}^{(i)},\omega^{(i)})-\chi^1_{g,d}(\boldsymbol{X}^{(i)},\tilde{\omega}_r^{(i)})\right),\\ \Delta^2_{g,d}(\alpha,\Omega)&=\sum_i\sum_{r=1}^R\alpha_r^{(i)}\left(\chi^2_{g,d}(\boldsymbol{X}^{(i)},\omega^{(i)})-\chi^2_{g,d}(\boldsymbol{X}^{(i)},\tilde{\omega}_r^{(i)})\right),\end{aligned} \tag{22}$$

ここで χ^0,χ^1,χ^2 はそれぞれ零次の統計量 (Occupancy)，一次の統計量，二次の統計量を収集する関数であり，以下の式で表わされる．

$$\begin{aligned}\chi^0_g(\boldsymbol{X}^{(i)},\omega)&=\sum_q\omega_q\sum_n I(q_n,g),\quad \chi^1_{g,d}(\boldsymbol{X}^{(i)},\omega)=\sum_q\omega_q\sum_n I(q_n,g)x^{(i)}_{n,d}\\ \chi^2_{g,d}(\boldsymbol{X}^{(i)},\omega)&=\sum_q\omega_q\sum_n I(q_n,g)\left(x^{(i)}_{n,d}\right)^2\end{aligned} \tag{23}$$

ここで $I(q_n,g)$ は時刻 n の隠れ変数 $q_n=(m_n,s_n)$ において，状態 s_n の混合要素 m_n が g 番目のガウス分布と等しい時に 1，そうでない時に 0 を取る指示関数である．これらの統計量は Forward-Backward アルゴリズムで効率的に計算が可能である．

混合重みベクトルの事後分布は以下のように示される．

$$P(\boldsymbol{\rho}_s|\alpha,\Omega)\propto\prod_m(\rho_{s,m})^{\hat{\phi}_{s,m}(\alpha,\Omega)} \tag{24}$$

ここで，$\hat{\phi}_{s,m}(\alpha,\Omega)$ は α,Ω が与えられた時の事後分布パラメタを示す関数であり，以下の式で示される．

$$\hat{\phi}_{s,m}(\alpha,\Omega)=\phi^{(0)}_{s,m}+\Delta^0_{G(s,m)}(\alpha,\Omega) \tag{25}$$

ここで $G(s,m)$ は状態 s，混合要素 m に対応する正規分布のインデックスを示す関数である．

2.6 目的関数の具体形

これらの事前分布を用いて，目的関数 (式 (14)) 内の期待値計算を展開すると以下を得る．

$$J(\alpha,\Omega)=\sum_i J_i^{\texttt{Slack}}(\alpha,\Omega)+\sum_g J_{g,d}^{\texttt{Gauss}}(\alpha,\Omega)+\sum_s J_s^{\texttt{Mix}}(\alpha,\Omega)+\text{constant}. \tag{26}$$

ここで，$J^{\texttt{Slack}}_{i,r}(\alpha,\Omega)$ はスラック変数 $\xi_r^{(i)}$ についての項で以下のように表わされる．

$$J^{\texttt{Slack}}_{i,r}(\alpha,\Omega)=\left(L^{(0)}-H(\omega^{(i)})+H(\tilde{\omega}_r^{(i)})\right)\sum_{r=1}^R\alpha_r^{(i)}+\log\left\{C^{(0)}-\sum_{r=1}^R\alpha_r^{(i)}\right\} \tag{27}$$

$J^{\texttt{Gauss}}_g(\alpha,\Omega)$ はガウス分布パラメタについての項で以下のように示される．

$$\begin{aligned}J^{\texttt{Gauss}}_g(\alpha,\Omega)=&\frac{\Delta^0_g(\alpha,\Omega)}{2}\log\{2\pi\}+\frac{\Delta^0_g(\alpha,\Omega)}{2}-\log\Gamma(\hat{\eta}_g(\alpha,\Omega))\\ &-\sum_{d=1}^D\hat{\eta}_g(\alpha,\Omega)\log\left\{\hat{R}_{g,d}(\alpha,\omega)\right\}.\end{aligned} \tag{28}$$

$J^{\texttt{Mix}}_s(\alpha,\Omega)$ は混合分布ベクトルに関する項で以下のように示される．

$$J^{\texttt{Mix}}_s(\alpha,\Omega)=\log\Gamma\left(\sum_m\phi^{(0)}_{s,m}+\Delta^0_{(s,m)}\right)-\sum_m\log\Gamma\left(\phi^{(0)}_{s,m}+\Delta^0_{(s,m)}\right). \tag{29}$$

これらの定義より，$\Omega,\tilde{\Omega}$ 内の各要素である離散確率分布 ω について，最適化や事後分布の算出に必要な量は，統計量 $\chi^0_g(X^{(i)},\omega),\chi^1_{g,d}(X^{(i)},\omega),\chi^2_{g,d}(X^{(i)},\omega)$ と，エントロピー $H(\omega)$ のみであることがわかる．後述の最適化法の実現ではこれらの要素を用いて ω を表現し，実際の ω の値の計算およびメモリ格納を避ける．

3. 並列動作可能な最適化アルゴリズム

本節では，前節で定義した双対目的関数 $J(\alpha,\Omega)$ を並列コンピューティング環境で最適化する手法について検討する．

3.1 α と Ω の交互最適化

α が与えられた上での Ω の最適化は，モデルが与えられた上での隠れ変数の事後確率推定にあたり，EM アルゴリズム (E-step) で効率良く並列計算することが可能であるが，α の最適化には EM アルゴリズムを適用することができない．そこで，Ω と α を異なる最適化アルゴリズムで交互に最適化することを考える．

以降，交互最適化における k 番目のステップにおける Ω の推定量を $\hat{\Omega}^{(k)}=\{\hat{\omega}^{(i,k)}\}$ と置き，α の推定量を $\hat{\alpha}^{(k)}\stackrel{\text{def}}{=}\{\hat{\alpha}_r^{(i,k)}|\forall i,1\le r\le R\}$ と書く．

k 番目のステップにおける Ω の推定量 ($\hat{\Omega}^{(k)}$) は一つ前のステップで推定された $\hat{\Omega}^{(k-1)}$ および $\hat{\alpha}^{(k-1)}$ を用いて，以下のように求めることができる．

$$\hat{\omega}^{(i,k)}=\underset{\omega}{\text{argmax}}\left\langle\tilde{\mathcal{L}}(\boldsymbol{X}^{(i)},A^{(i)};\Theta,\omega)\right\rangle_{P(\Theta|\hat{\alpha}^{(k-1)},\hat{\Omega}^{(k-1)})} \tag{30}$$

この最適化は各トレーニングデータ i に対して独立な形となっているため，トレーニングデータ毎に並列に実行することができる．さらに，この最適化は Forward-Backward アル

Vol.2010-SLP-80 No.8
2010/2/13

ゴリズムを用いることで厳密に求めることが可能であり，効率的に並列計算することができる．

k 番目のステップにおける $\hat{\alpha}^{(k)}$ は得られた $\hat{\Omega}^{(k)}$ を用いて，以下の最適化問題を解くことで得ることができる．

$$\hat{\alpha}^{(k)} = \underset{\alpha}{\mathrm{argmax}} J(\alpha, \hat{\Omega}^{(k)}) \tag{31}$$

この最適化は $\hat{\Omega}^{(k)}$ の場合と異なり，解析的に求めることができない．そこで，並列化可能な反復法を用いて数値的に解くことを試みる．

本研究では，この最適化に Rprop に基づく方法を用いた．Rprop は収束の速い勾配法の一つであり，並列化が容易である．また，多くの先行研究で利用されており，様々な目的関数に対して安定した性能を示すことが知られている．Rprop も反復法であるため，k とは別にステップ番号 t を導入し，t ステップ目における $\hat{\alpha}^{(k)}$ を $\hat{\alpha}^{(k,t)} \stackrel{\mathrm{def}}{=} \{\hat{\alpha}_r^{(i,k,t)} | \forall i, 1 \le r \le R\}$ で表わすと，Rprop の (k,t) ステップ目における $\hat{\alpha}_r^{(i,k,t)}$ は以下のように表わされる．

$$\begin{aligned} \hat{\alpha}_r^{(i,k,t)} =& \hat{\alpha}_r^{(i,k,t-1)} + \nabla_r^{(i,k,t-1)} \cdot \varsigma^{(i,k,t-1)} \\ \varsigma^{(i,k,t)} \stackrel{\mathrm{def}}{=}& \mathrm{sign}\left(\frac{\delta}{\delta \alpha_r^{(i)}} J\left(\hat{\alpha}^{(i,k,t)}, \hat{\Omega}^{(k)}\right)\right) \end{aligned} \tag{32}$$

ここで $\mathrm{sign}(x)$ は x が正の時に $+1$，負の時に -1，ゼロの時に 0 を取る関数である．また，$\nabla_r^{(i,k,t)}$ は目的関数の $\hat{\alpha}^{(k,t)}$ における勾配の符号 $\varsigma^{(i,k,t)}$ によって決定される移動量であり，一定の弾性力を持つ値として，以下のように定義される．

$$\nabla_r^{(i,k,t)} = \begin{cases} \nabla^{\mathtt{INIT}} & t = 1 \\ \min\{\nabla^{\mathtt{MAX}}, \nu^+ \cdot \nabla_r^{(i,k,t-1)}\} & \varsigma^{(i,k,t-1)} \cdot \varsigma^{(i,k,t)} > 0 \\ \max\{\nabla^{\mathtt{MIN}}, \nu^- \cdot \nabla_r^{(i,k,t-1)}\} & \varsigma^{(i,k,t-1)} \cdot \varsigma^{(i,k,t)} < 0 \\ \nabla_r^{(i,k,t-1)} & \varsigma^{(i,k,t-1)} \cdot \varsigma^{(i,k,t)} = 0 \end{cases} \tag{33}$$

ここで，$\nabla^{\mathtt{INIT}}, \nabla^{\mathtt{MIN}}, \nabla^{\mathtt{MAX}}, \nu^+, \nu^-$ は最適化のチューニングパラメタであり，最適化の収束速度を決定する．この更新則による α の最適化は i および r ごとに独立して行なうことができるため，並列化が容易である．

本研究で提案した最小相対エントロピー識別学習では，最適化の変数であるラグランジュ未定乗数の非負制約があり，かつ，目的関数の収束領域が限られているという問題があり，最急降下法のような勾配法の適用は，勾配が発散してしまう領域があるという点で難しい．Rprop は勾配を元にした最適化技法であるが，勾配ベクトル内の各要素の符号のみに着目した最適化を行なうため，目的関数，およびその微分値が発散してしまうような場合でも，特殊な処理を挿入することなく最適化を行なうことができる．

3.2 切除平面法に基づくサブセット $\tilde{\Omega}^{(i)}$ の決定

制約のサブセット $\tilde{\Omega}^{(i)}$ の決定には切除平面法に基づく方法を用いた．切除平面法は最初，整数計画問題の解法として導入されたが，Tsochantaridis 等によって制約の数が無限にある SVM (識別対象となるクラス数が無限であるマルチクラス SVM) の解法としても用いられてきた [8]．また，切除平面法は従来の MRED の実装にも用いられていたが，従来の実装では，最終的に N-Best 仮説集合を得るという形に帰着されていた．本稿で提案する手法は統計量に相当する $\tilde{\omega}_r^{(i)}$ をラティスから得ようとするもので，ラティス上での実装に関して必要なものである．

表 1 Rprop を用いたラティス型 MRED の最適化
Table 1 The lattice-based MRED optimization method based on Rprop

```
1: k ← 1
2: loop
3:    正解ラティスについての χ̂^(i) (式 (35)) を全ての i について計算 (E-step に相当)
4:    不正解ラティスについての χ̃_r^(i) (式 (35)) を全ての i, r について計算
5:    t ← 1
6:    α_r^(i,k,0) ← 0.0   (∀i, ∀r)
7:    ∇_r^(i,k,0) ← ∇_INIT   (∀i, ∀r)
8:    loop
9:       目的関数の微分の符号 ς_r^(i,k,t) を全ての i, r について計算
10:      ∇_r^(i,k,t) を式 (33) に従って決定
11:      α_r^(i,k,t) ← max{0.0, min{C, α_r^(i,k,t-1) + ς^(i,k,t) ∇_r^(i,k,t)}}
12:      得られた α_r^(i,k,t) を用いて累積差分統計量 Δ^0, Δ^1, Δ^2 (式 (22)) を計算
13:   end loop
14: end loop
```

前節で提案した制約付き最適化のサブセットを用いた近似は $R \to \infty$ の極限で厳密解となる．一般に，制約付き最適化問題の制約の多くは冗長であることが知られており，一つの制約が満たされれば，他の制約も満たされることが多い．KKT 条件から，ある制約が満たされた場合，その制約に対応するラグランジュ未定乗数の最適解が 0 になることが示されている [9]．切除平面法では，他の制約を従属させるようなクリティカルな制約を順次追加していくことで，ラグランジュ未定乗数が 0 になるような制約を考慮せずに最適化を行なう．特に，Tsochantaridis らの方法では，推定中にある解の仮説を用いて，最も充足させるのが難しい制約を順次追加することで切除平面法を効率的に実現してきた．

この考えを本手法に適用し，交互最適化中の解仮説 $P^{(k-1)}(\Theta, \xi) \stackrel{\mathrm{def}}{=} P(\Theta, \xi | \hat{\alpha}^{(k-1)}, \Omega^{(k-1)})$ を用いて，最も制約を充足しにくい $\tilde{\omega}_k^{(i)}$ を順次追加していくことを考えると，$\tilde{\omega}_k^{(i)}$ は以下のように定義される．

$$\begin{aligned} \tilde{\omega}_k^{(i)} =& \underset{\tilde{\omega}}{\mathrm{argmin}} \left\langle \tilde{\mathcal{L}}(\boldsymbol{X}^{(i)}, A^{(i)}; \Theta, \omega) - \tilde{\mathcal{L}}(\boldsymbol{X}^{(i)}, \tilde{A}^{(i)}; \Theta, \tilde{\omega}) - \xi^{(i)} \right\rangle_{P^{(k-1)}(\Theta,\xi)} \\ =& \underset{\tilde{\omega}}{\mathrm{argmax}} \left\langle \tilde{\mathcal{L}}(\boldsymbol{X}^{(i)}, \tilde{A}^{(i)}; \Theta, \tilde{\omega}) \right\rangle_{P^{(k-1)}(\Theta,\xi)} \end{aligned} \tag{34}$$

この最適化の解 $\tilde{\omega}_r^{(i)}$ は EM アルゴリズムの E-step を不正解ラティス $\tilde{A}^{(i)}$ に対して適用した際に得られる統計量で示される．これを元の最適化に組み込むと，交互最適化が 1 ステップ進むごとに R が一つ増え，それに対応する $\alpha_r^{(i)}$ が最適化対象として追加される最適化プロセスを得ることができる．

以上全ての処理を疑似コードで示したものを表 1 に示す．先述したように，アルゴリズム中では $\hat{\omega}^{(i)}$ および $\tilde{\omega}_r^{(i)}$ の直接表現は避け，これらによって決定される以下のような統計量を用いている．

$$\begin{aligned} \hat{\chi}^{(i)} \stackrel{\mathrm{def}}{=}& \left\{ H(\hat{\omega}^{(i)}), \chi_g^0(\boldsymbol{X}^{(i)}, \hat{\omega}^{(i)}) \chi_{g,d}^1(\boldsymbol{X}^{(i)}, \hat{\omega}^{(i)}), \chi_{g,d}^2(\boldsymbol{X}^{(i)}, \hat{\omega}^{(i)}) | \forall g, \forall d \right\}, \\ \tilde{\chi}_r^{(i)} \stackrel{\mathrm{def}}{=}& \left\{ H(\tilde{\omega}_r^{(i)}), \chi_g^0(\boldsymbol{X}^{(i)}, \tilde{\omega}_r^{(i)}), \chi_{g,d}^1(\boldsymbol{X}^{(i)}, \tilde{\omega}_r^{(i)}), \chi_{g,d}^2(\boldsymbol{X}^{(i)}, \tilde{\omega}_r^{(i)}) | \forall g, \forall d \right\}. \end{aligned} \tag{35}$$

一般に ω_q を直接メモリ上に格納することは困難であるため，こうした間接的な表現を用いることは必須であると考えられる．

表 2 比較した手法の音素正解精度
Table 2 Phoneme Accuracies of the compared methods.

# Mix. Comp.	MLE	MRED (1-best)	MRED (Lattice)
1	58.0	59.7	61.2
2	61.7	N/A	64.5
4	64.4	N/A	66.2
8	67.0	N/A	68.0

表 3 計算に利用したスレッド数と計算性能の関係
Table 3 Normalized computational speed as a function of the numbers of threads used

# Threads	1	2	4	8	16	32
Forward-Backward	1.0	1.74	3.36	5.81	10.17	13.31
Rprop	1.0	1.70	3.30	5.78	8.51	9.04

4. 実験と考察

提案法の有効性を音素バイグラムモデルを用いた連続音素認識タスクで評価した．

テストセット及びトレーニングセットは TIMIT データベースから構築した．トレーニングセットの総発話数は 3,696 発話，テストセットの総発話数は 192 発話，HMM の構造は文脈非依存の 3 状態 Left-to-Right 型音素 HMM とした．HMM の総数すなわち識別対象となる音素数は 48，評価時はこれを文献 [11] の方法で 39 個の代表音素に縮約し，代表音素の正解精度を評価した．

n 番目の正規分布パラメタの事前分布パラメタは，I-smoothing と同様に EM アルゴリズムによって得られた統計量を用い，以下のように定義した

$$\gamma_g^{(0)} = \tilde{\chi}_g^0, \quad \mu_{g,d}^{(0)} = \gamma_g^{(0)} \tilde{\chi}_{g,d}^1, \quad \eta_g^{(0)} = \frac{1}{2} \sum_i \tilde{\chi}_g^0, \quad R_{g,d}^{(0)} = \tilde{\chi}_{g,d}^2, \quad \phi_{s,m}^{(0)} = \tilde{\chi}_{G(s,m)}^0. \tag{36}$$

ここで $\tilde{\chi}_g^0, \tilde{\chi}_{g,d}^1, \tilde{\chi}_{g,d}^2$ は最尤推定によって得た CD-HMM を用いて EM アルゴリズムの E-Step を実行することによって収集した g 番目のガウス分布に関する零次／一次／二次の統計量の d 次元目である．また，スラックのハイパーパラメタとしては以下を用いた．

$$C^{(0)} = 1, \quad L^{(0)} = 0 \tag{37}$$

Rprop に利用される定数としては以下を用いた．

$$\nabla^{\mathtt{INIT}} = C^{(0)}/4, \quad \nabla^{\mathtt{MIN}} = 0.0, \quad \nabla^{\mathtt{MAX}} = C^{(0)}/4, \quad \nu^+ = 1.2, \quad \nu^- = 0.5 \tag{38}$$

提案法は α に対して凹であるため，この設定は最適化の結果には全く影響しない．

比較対象として最尤推定 (MLE) によって学習した HMM を利用した．表 2 に提案法および最尤推定法の音素正解精度を示す，また，ラティスを用いずに 1-Best シーケンスを認識仮説として学習した MRED 法の結果を既報の論文 [4] より転記する．

結果より，MRED が識別学習法として最尤推定より高い精度を達成できていること，およびラティスを用いることでさらに精度を向上させることが可能であることが確認できた．また，提案法は並列コンピューティング環境で効率的に動作させることができる．グリッドコンピュータの利用によって，従来よりさらに大きなモデル，具体的には混合分布を含む CD-HMM の学習も可能になった．表 3 に，シングルスレッド実行時の性能を 1 とした際の計算性能比と利用した計算スレッドの数の関係を示す．表中の “Forward-Backward” は提案法中に含まれる Ω-最適化を実行する部分の並列化性能であり，Forward-Backward 法が比較的並列計算しやすい問題であることから，この並列化効率は並列計算環境そのものの性能限界であると考えることができる．今回導入した α-最適化 (Rprop) においては，8 スレッドまでは Forward-Backward とほぼ同等の並列化性能を達成することができた．16 スレッド以上での性能向上は大きくなかったが，これは α の最適化において各計算ノードが必要とする情報である，事前分布および累積差分統計量の送信コストによるものだと考えられる．これらの送信コストを含む計算ノード間通信コストを最適化するため，より高度な分散ストレージ技術との統合が望まれる．

5. まとめと今後の課題

本稿では，連続分布型隠れマルコフモデルの最小相対エントロピー識別を用いた識別学習において，データセットの量に対するスケーラビリティを向上させるための手段として，その学習アルゴリズムに仮説候補のラティスによる表現と勾配法による最適化を導入した．仮説候補のラティスによる表現は対立候補を複数用意することに起因する冗長性をなくすことに貢献し，勾配法による最適化は並列処理を行なう際の効率性の向上に貢献する．これらの導入によって，従来より規模の大きな実験を行なう際のソフトウェア基盤が完成した．

最小相対エントロピー識別の実装法に関する今後の課題としては，より効率の高い，並列最適化法を導入することが挙げられる．また，最小相対エントロピーの理論に関する今後の課題としては，MPE のように認識誤りの仮説ひとつひとつに対して異なるエラー尺度を与える手法に関する検討と，高性能を達成する事前分布に関する検討がある．

参 考 文 献

1) E. McDermott, S. Katagiri, “String-Level MCE for Continuous Phoneme Recognition,” Proc. EUROSPEECH-1997, pp. 123–126, 1997.
2) P.C. Woodland, D. Povey, “Large Scale Discriminative Training of Hiddel Markov Models for Speech Recognition,” Computer Speech and Language, Vol. 16, pp. 25–47, 2002.
3) T. Jebara, “Machine Learning: Discriminative and Generative,” Kluwer Academic Publishers, 2004.
4) 久保，渡部，中村，白井，“最小相対エントロピー基準によるパラメタ分布の正則化を用いた連続分布 HMM の識別学習，” 日本音響学会春季全国大会講演論文集, 2009 年 3 月.
5) L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer, “Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition,” Proc. ICASSP’86, pp. 49-52, 1986.
6) D. Povey, P.C. Woodland, “Minimum Phone Error and I-Smoothing for Improved Discriminative Training,” Proc. ICASSP-2002, Vol. 1, pp. 105–108, 2002.
7) C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, M. Mohri, “OpenFst: A General and Efficient Weighted Finite-State Transducer Library – (Extended Abstract of an Invited Talk), ” Lecture Notes in Cmputer Science, Vol. 4783, pp. 11–23, 2007.
8) I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, “Large Margin Methods for Structured and Interdependent Output Variables,” Journal of Machine Learning Research, Vol. 6, pp. 1453–1484, 2005.
9) B. Schölkopf, A. Smola, “Learning with Kernels,” pp. 165–175, MIT Press, 2002.
10) C. Igel, M. Hüsken, “Improving the Rprop Learning Algorithm,” Proc. of the 2nd International ICSC Symposium on Neural Computation (NC–2000), pp.115–121, 2000.
11) K-F. Lee, H. Hon, “Speaker-Independent Phone Recognition Using Hidden Markov Models,” IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 11, 1989.